DOI: 10.18137/RNU.V9187.23.04.P.176 УДК 004.912

Э.А. Челышев, М.В. Раскатова, А.С. Маковец

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО КВАЗИРЕФЕРИРОВАНИЯ ТЕКСТОВ

Аннотация. Приводится постановка задачи автоматического квазиреферирования текстов, а также подробно рассматриваются такие алгоритмы автоматического квазиреферирования текстов, как алгоритм Луна, латентный семантический анализ, TextRank и LexRank. Выполнена оценка информационной полноты для набора рефератов, сгенерированных при помощи указанных алгоритмов, с использованием метрик информационной близости: метрики, основанной на расстоянии Дженсена – Шеннона, и косинусного сходства, примененных к векторным представлениям исходного текста и полученных рефератов. Проведен статистический анализ полученных результатов.

Ключевые слова: реферат, реферирование, квазиреферирование, расстояние Дженсена – Шеннона, косинусное сходство.

E.A. Chelyshev, M.V. Raskatova, A.S. Makovets

COMPARATIVE ANALYSIS OF AUTOMATIC TEXT QUASI-SUMMARIZATION ALGORITHMS

Abstract. The article presents the formulation of the problem of automatic text quasi-summarization and also discusses in detail such algorithms of automatic text quasi-summarization as the Luhn's algorithm, Latent Semantic Analysis, TextRank and LexRank. The information completeness was evaluated for a set of abstracts generated using these algorithms. The information completeness evaluation was performed using information proximity metrics: a metric based on the Jensen – Shannon distance and cosine similarity applied to vector representations of the source text and the resulting abstracts. A statistical analysis of the obtained results was carried out.

Keywords: summary, summarization, quasi-summarization, Jensen – Shannon divergence, cosine similarity.

Введение

Человечество за свою историю накопило огромные массивы текстовых данных. Современный человек вынужден постоянно взаимодействовать с различными текстами, в том числе достаточно объемными, на ознакомление с которыми могут потребоваться существенные временные затраты. Именно по этой причине возникает необходимость в реферировании текстов, то есть в подготовке нового текста существенно меньших объемов, чем исходный текст, но при этом содержащего основные факты и сведения исходного. Получаемый в процессе реферирования текст называется рефератом.

Реферирование текстов может быть выполнено автоматически. Исследования в области автоматического реферирования текстов ведутся с 1958 года, когда американский ученый немецкого происхождения Ханс Петер Лун представил научному сообществу первую известную работу, посвященную данной проблематике [1].

Среди алгоритмов автоматического реферирования текстов по типу получаемого реферата выделяют экстрагирующие, абстрагирующие и гибридные алгоритмы.

При использовании экстрагирующих алгоритмов реферат составляется из наиболее значимых предложений исходного текста, которые при этом не изменяются. Правила вы-

Челышев Эдуард Артурович
аспирант, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: машинное обучение по прецедентам, искусственные нейронные сети, машинная обработка текстов на естественном языке, C++. Автор более 20 опубликованных научных работ. SPIN-код: 5357-7604, AuthorID: 1088395.
Электронный адрес: chel.ed@yandex.ru

Раскатова Марина Викторовна
кандидат технических наук, доцент кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: разработка программного обеспечения, информационные системы. Автор более 60 опубликованных научных работ. AuthorID: 609945.
Электронный адрес: RaskatovaMV@mpei.ru

Маковец Антон Сергеевич
ассистент кафедры вычислительных машин, систем и сетей, Национальный исследовательский университет «МЭИ», Москва. Сфера научных интересов: языки и методы программирования, цифровая схемотехника.
Электронный адрес: MakovetsAnS@mpei.ru

числения значимости предложения определяются конкретным алгоритмом [2]. Абстрагирующий подход, в свою очередь, характерен тем, что при его использовании генерируется новый текст, меньший по объему, чем исходный, но при этом удовлетворяющий всем правилам языка, на котором был составлен исходный текст. Гибридный подход комбинирует в себе приемы как экстрагирующего, так и абстрагирующего подходов [3].

Реферирование с использованием экстрагирующего подхода называется также квазиреферированием, а получаемый таким образом реферат – **квазирефератом** [4].

В дальнейшем в данной работе используются следующие термины:

терм – минимальная смысловая единица в тексте (то же самое, что слово);

документ – смысловая единица текста, включающая в себя набор термов, объединенных в предложения;

корпус – совокупность документов;

словарь – совокупность всех рассматриваемых термов некоторого текстового корпуса.

Алгоритмы автоматического квазиреферирования текстов

В данной работе рассмотрены четыре алгоритма автоматического квазиреферирования текстов: алгоритм Луна, латентный семантический анализ, TextRank и LexRank.

Алгоритм Луна исторически является первым известным алгоритмом автоматического квазиреферирования текстов. В современном понимании данный алгоритм состоит из последовательно выполняемых этапов: определение значимых термов, вычисление значимостей предложений, отбор необходимого числа предложений из наиболее значимых [1].

Этап определения значимых термов состоит из следующих шагов:

1) формирование словаря термов исходного текста; на данном этапе исходный текст разбивается на отдельные термы;

2) нормализация термов словаря, при которой производится приведение всех термов словаря к некоторой нормальной форме (например, словарной); нормализация термов

выполняется при помощи методов морфологического анализа, таких как стемминг и лемматизация [5];

3) исключение из словаря стоп-слов, не несущих существенной семантической нагрузки: предлогов, союзов, частиц.

В результате шагов 2 и 3 размер словаря сокращается, что положительно сказывается на качестве и скорости работы алгоритма.

Значимыми признаются термы, чья частота встречаемости в исходном тексте выше наперед заданного порогового значения.

Вычисление значимостей предложений происходит по следующему алгоритму: на вход подается реферируемый текст $T=(s_1,s_2,\ldots,s_n)$. Для каждого предложения s_i определяется набор последовательностей термов данного предложения $\Omega_i=\{\omega_1,\omega_2,\ldots,\omega_k\}$, каждый элемент которых удовлетворяет следующему условию: последовательность термов ω_j входит в формируемый набор тогда и только тогда, когда она начинается и заканчивается значимым термом, а максимальное количество подряд идущих незначащих термов не более наперед заданного числа (как правило, 4–5 термов).

Для каждой последовательности термов ω_j вычисляется значимость $\Xi(\omega_j)$ по формуле

$$\Xi(\omega_j)=\frac{N^2_{знач}(\omega_j)}{|\omega_j|}, \tag{1}$$

где $N_{знач}(\omega_j)$ – количество значащих слов в последовательности T; $|\omega_j|$ – количество элементов в последовательности T.

Значимость предложения s_i при этом определяется как максимум из значимостей последовательностей данного предложения по формуле

$$\Xi(s_i)=\max\{\Xi(\omega):\omega\in\Omega_i\}, \tag{2}$$

Латентный семантический анализ (англ. Latent Semantic Analysis, LSA) применяется для поиска семантических отношений в тексте и использует сингулярное разложение матриц.

Матрица A, составленная таким образом, что элемент A_{ij} несет значение встречаемости терма t_j в документе d_i, называется **термдокументной**. Если условиться, что предложения исходного текста являются документами, то такая матрица будет представлять реферируемый текст. Затем над данной матрицей проводится сингулярное разложение. В реферат включается необходимое число предложений, которым соответствуют наибольшие сингулярные значения [6].

Алгоритм TextRank является графовым алгоритмом автоматического квазиреферирования текстов и основан на алгоритме ранжирования веб-страниц PageRank [7]. Строится полносвязный граф, вершинам которого соответствуют предложения исходного текста. Вес ребра w_{ij} между вершинами V_i и V_j взвешивается по формуле

$$w_{ij}=\frac{|t_k : t_k\in P_i \,u\, t_k\in P_j|}{\log(|P_i|)+\log(|P_j|)}, \tag{3}$$

где P_i – предложение; t_k – отдельный терм; $|P_i|$ – длина предложения (количество термов в нем).

Затем для каждой вершины (предложения исходного текста) определяются веса в соответствии с формулой, аналогичной используемой в алгоритме PageRank:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ij}}{\left|Out(V_j)\right|} S(V_i). \quad (4)$$

Алгоритм LexRank также является графовым алгоритмом автоматического квазиреферирования текстов, однако имеет ряд отличий от TextRank. Алгоритм LexRank требует предварительной генерации векторных представлений предложений исходного текста с использованием какого-либо метода векторизации текста. При этом вес ребра, соединяющего *i*-ю и *j*-ю вершины, определяется с использованием метрики, определенной на векторном пространстве (например, косинусное расстояние). Затем из графа исключаются ребра, имеющие вес, не превышающий некоторого наперед заданного порогового значения. Значимыми признаются предложения, для которых соответствующие вершины имеют наибольшее количество смежных вершин, а также вершин-соседей второго порядка [8].

Оценка информационной полноты реферата

Качество реферата – комплексный показатель, учитывающий его степень сжатия относительно исходного текста, качество языка и информационную полноту реферата, характеризующую, насколько полно сгенерированный реферат отражает содержание исходного текста. В работе проводилась именно оценка последнего из перечисленных показателей качества.

В данной работе для оценки информационной полноты квазиреферата использовались метрики информационной близости: метрика, основанная на расстоянии Дженсена – Шеннона, определяемая по формуле, приведенной ниже, и косинусное сходство [9; 10]:

$$\overline{D_{JS}}(P||Q) = 1 - D_{JS}(P||Q), \quad (5)$$

где $D_{JS}(P||Q)$ – значение расстояния Дженсена – Шеннона для вероятностных распределений P и Q.

Данные метрики применялись к векторным представлениям исходного текста и текста сгенерированного квазиреферата. Значения данных метрик нужно трактовать следующим образом: чем ближе их значение к нулю, тем менее схожими являются тексты, следовательно, тем хуже квазиреферат соответствует исходному тексту; чем ближе значения метрик к единице, тем больше оба текста соответствуют друг другу.

В данной работе для выполнения векторизации использовались два метода: частотная векторизация и векторизация методом TF-IDF [11]. При этом метрика, основанная на расстоянии Дженсена – Шеннона, применялась исключительно в рамках частотной векторизации, так как расстояние Дженсена – Шеннона вычисляется только для вероятностных распределений. Все эксперименты, обработка данных и вычисление метрик проводились с использованием языка программирования Python и его библиотек.

Для оценки информационной полноты рассматриваемых методов использовался корпус русскоязычных новостных статей веб-портала LENTA.RU [12]. Из имевшегося корпуса были отобраны 146 текстов длиной от 40 до 50 предложений, относящиеся к следующим рубрикам: «Бизнес», «Культура», «Наука и техника», «Спорт», «Экономика». Для каждого из документов был получен квазиреферат с использованием каждого из рассматриваемых алгоритмов, затем проведена оценка информационной полноты каждого из квазирефератов, а также квазиреферата, полученного случайным выбором предложений исходного текста.

Значения метрик информационной близости приведены в Таблицах 1–3 отдельно по каждой рубрике, а также по целому набору. Средние значения метрик визуализированы в виде гистограмм на Рисунках 1–3.

Таблица 1

Значения метрики, основанной на расстоянии Дженсена – Шеннона, для частотной векторизации текстов

Текст	Методы				
	Метод Луна	LexRank	TextRank	LSA	Случайный выбор
Бизнес	0,523	0,441	0,544	0,489	0,479
Культура	0,443	0,403	0,474	0,476	0,385
Наука и техника	0,515	0,478	0,555	0,560	0,445
Спорт	0,497	0,458	0,520	0,511	0,370
Экономика	0,440	0,398	0,457	0,425	0,380
Среднее значение	0,473	0,431	0,500	0,491	0,391

Таблица 2

Значения косинусного сходства для частотной векторизации текстов

Текст	Методы				
	Метод Луна	LexRank	TextRank	LSA	Случайный выбор
Бизнес	0,856	0,783	0,851	0,789	0,796
Культура	0,766	0,724	0,769	0,755	0,677
Наука и техника	0,825	0,786	0,837	0,817	0,743
Спорт	0,808	0,776	0,818	0,785	0,656
Экономика	0,795	0,764	0,785	0,718	0,693
Среднее значение	0,796	0,757	0,800	0,769	0,688

Таблица 3

Значения косинусного сходства для векторизации методом TF-IDF

Текст	Методы				
	Метод Луна	LexRank	TextRank	LSA	Случайный выбор
Бизнес	0,845	0,746	0,852	0,720	0,715
Культура	0,756	0,726	0,765	0,717	0,643
Наука и техника	0,802	0,753	0,799	0,796	0,709
Спорт	0,768	0,758	0,774	0,726	0,664
Экономика	0,761	0,734	0,746	0,663	0,649
Среднее значение	0,771	0,742	0,774	0,722	0,660

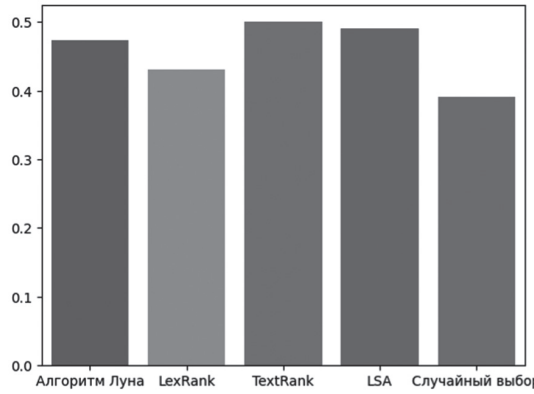

Рисунок 1. Средние значения метрики, основанной на расстоянии Дженсена – Шеннона, для частотной векторизации

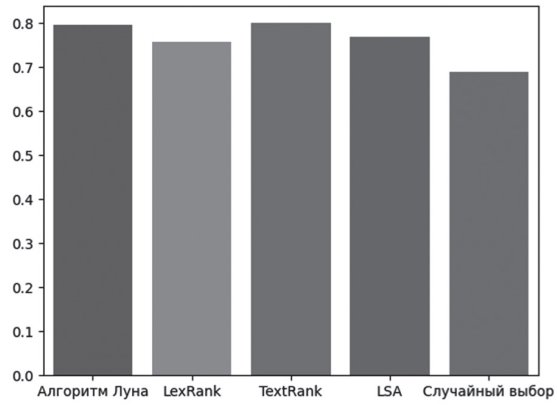

Рисунок 2. Средние значения косинусного сходства для частотной векторизации

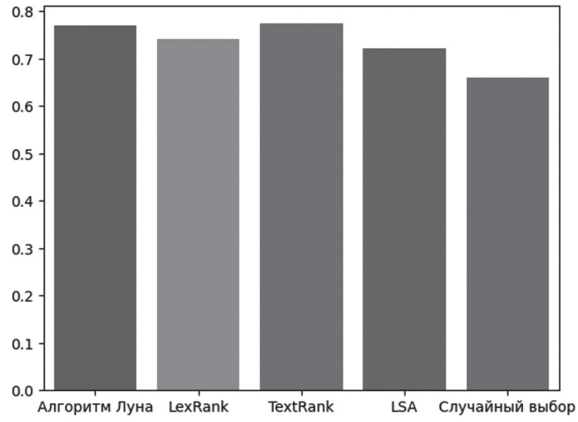

Рисунок 3. Средние значения косинусного сходства для векторизации методом TF-IDF

В Таблице 4 приведены значения коэффициента корреляции Пирсона между полученными массивами значений метрик для различных экспериментов.

Таблица 4

Значения коэффициента корреляции Пирсона для различных метрик

Метрики	$\overline{D_{JS}}$ (част. векст.)	*CS* (част. векст.)	*CS* (TF-IDF)
$\overline{D_{JS}}$ (част. векст.)	1	0,903773	0,776802
CS (част. векст.)	0,903773	1	0,968129
CS (TF-IDF)	0,776802	0,968129	1

Заключение

Полученные в ходе экспериментов значения метрик информационной близости обнаруживают высокие значения коэффициента корреляции Пирсона между собой. Особенно заметно коррелируют со всеми остальными значениями результаты, полученные с использованием косинусного сходства для частотной векторизации.

Таким образом, можно говорить о наличии зависимости между результатами, полученными для различных метрик информационной близости и различных методов векторизации текста.

Во всех рассмотренных случаях результаты, полученные для алгоритмов автоматического квазиреферирования текстов, оказались существенно лучше аналогичных показателей для квазирефератов, получаемых случайной выборкой предложений исходного текста. При этом и в случае частотной векторизации, и при использовании метода TF-IDF лучше себя показали такие алгоритмы автоматического реферирования, как алгоритм Луна и TextRank, наихудшие результаты – у алгоритмов LexRank и латентно-семантического анализа.

Литература

1. *Luhn H.* The automatic creation of literature abstracts // IBM Journal of Research and Development. 1958.Vol. 2. Pp. 159–165. DOI: 10.1147/rd.22.0159

2. *Бакиева А.М., Батура Т.В., Федотов А.М.* Методы и системы автоматического реферирования текста // Вычислительные и информационные технологии в науке, технике и образовании (CITech-2015). Алматы, 24–27 сентября 2015 г. Ч. 1. Алматы : Казахский национальный университет имени Аль-Фараби, 2015. С. 263–274. EDN UJOVTL.

3. *Батура Т.В., Бакиева А.М.* Методы и системы автоматического реферирования текстов. Новосибирск : Новосибирский национальный исследовательский государственный университет, 2019. 110 с. ISBN 978-5-4437-0974-1. EDN VLYLOR.

4. *Полицына Е.В., Полицын С. А., Касаткина А. О.* Создание интегрального алгоритма и инструментов автоматического реферирования текстов на русском языке // Информационные технологии. 2020. Т. 26. № 1. С. 30–38. EDN KCJZGM. DOI: 10.17587/it.26.30-38

5. *Челышев Э.А., Оцоков Ш.А., Раскатова М.В., Щеголев П.* Сравнение методов классификации русскоязычных новостных текстов с использованием алгоритмов машинного обучения // Вестник кибернетики. 2022. № 1 (45). С. 63–71. EDN VHTYBB. DOI: 10.34822/1999-7604-2022-1-63-71

6. *Kumar Y.J., Goh O.S., Basiron H.* A review on automatic text summarization approaches // Journal of Computer Science. 2016. Vol. 12. No. 4. Pp. 178–190. DOI: 10.3844/jcssp.2016.178.190

7. *Louis A., Nenkova A.* Automatic Summary Evaluation without Human Models // Theory and Applications of Categories. 2008. URL: https://tac.nist.gov/publications/2008/additional.papers/Penn.proceedings.pdf (дата обращения: 12.04.2023).

8. *Крылов В.С.* Автоматическое аннотирование текста с помощью R-пакета LexRank // Информационно-компьютерные технологии в экономике, образовании и социальной сфере. 2022. № 3 (37). С. 73–81. EDN TSCNBZ.

9. *Lin C.-Y., Cao G., Gao J., Nie J.-Y.* An Information-Theoretic Approach to Automatic Evaluation of Summaries // Proceedings of the Human Language Technology Conference of the NAACL. Main Conference, New York City, USA, 2006. Pp. 463–470. URL: https://aclanthology.org/N06-1059/# (дата обращения: 12.04.2023).

10. *Manning C.D., Raghavan P., Schütze H.* Introduction to Information Retrieval. Cambridge, England : Cambridge University Press, 2008. 482 p. ISBN 0521865719.

11. *Раскатова М.В., Челышев Э.А.* Векторизация текстов в задачах обработки естественного языка: история и развитие // Современное программирование : Материалы IV Международной научно-практической конференции, Нижневартовск, 08 декабря 2021 г. / Под общ. ред. Т.Б. Казиахмедова. Нижневартовск : Нижневартовский государственный университет, 2022. С. 284–288. EDN BZQQVZ. DOI: 10.36906/AP-2022/47

12. News dataset from Lenta.Ru. Corpus of Russian news articles collected from Lenta.Ru // Kaggle.com. URL: https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta (дата обращения: 12.04.2023).

References

1. Luhn H. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*. Vol. 2. Pp. 159–165. DOI: 10.1147/rd.22.0159

2. Bakieva A.M., Batura T.V., Fedotov A.M. (2015) Methods and systems for automatic text summarization. In: *Vychislitel'nye i informatsionnye tekhnologii v nauke, tekhnike i obrazovanii (CITech-2015)*. Almaty, September 24–27, 2015. Part 1. Almaty : Al-Farabi Kazakh National University. Pp. 263–274. (In Russian).

3. Batura T.V., Bakieva A.M. (2019) *Metody i sistemy avtomaticheskogo referirovaniya tekstov* [Methods and systems for automatic text summarization]. Novosibirsk : Novosibirsk State University Publ. 110 p. ISBN 978-5-4437-0974-1. (In Russian).

4. Politsyna E.V., Politsyn S.A., Kasatkina A.O. (2020) Development of integrated algorithm and tools of automatic summarization for texts in the Russian language. *Informatsionnye tekhnologii* [Information Technology]. Vol. 26. No. 1. Pp. 30–38. DOI: 10.17587/it.26.30-38 (In Russian).

5. Chelyshev E.A., Otsokov Sh.A., Raskatova M.V., Shchegolev P. (2022) Comparison of methods for classifying Russian-language news texts using machine learning algorithms. *Vestnik kibernetiki [Proceedings in Cybernetics]*. No. 1. Pp. 63–71. DOI: 10.34822/1999-7604-2022-1-63-71 (In Russian).

6. Kumar Y.J., Goh O.S., Basiron H. (2016) A review on automatic text summarization approaches. *Journal of Computer Science*. Vol. 12, No. 4. Pp. 178–190. DOI: 10.3844/jcssp.2016.178.190

7. Louis A., Nenkova A. (2008) Automatic Summary Evaluation without Human Models. *Theory and Applications of Categories*. URL: https://tac.nist.gov/publications/2008/additional.papers/Penn.proceedings.pdf (accessed 12.04.2023).

8. Krylov V.S. (2022) Automatic text annotation using the LexRank R package. *Informatsionno-komp'yuternye tekhnologii v ekonomike, obrazovanii i sotsial'noy sfere* [Information and computer technologies in the economy, education and social sphere]. No. 3. Pp. 73–81. (In Russian).
9. Lin C.-Y., Cao G., Gao J., Nie J.-Y. (2006) An Information-Theoretic Approach to Automatic Evaluation of Summaries. In: Proceedings of the Human Language Technology Conference of the NAACL. Main Conference, New York City, USA, June 2006. Pp. 463–470. URL: https://aclanthology.org/N06-1059/# (accessed 12.04.2023).
10. Manning C.D., Raghavan P., Schütze H. (2008) *Introduction to Information Retrieval.* Cambridge, England: Cambridge University Press, 2008. 482 p. ISBN 0521865719.
11. Raskatova M.V., Chelyshev E.A. (2022) Vectorization of texts in natural language processing problems: History and development. In: Kaziakhmedov T.B. (Ed) *Sovremennoe programmirovanie* [Modern programming] : Proc. IV Int. Sci. and Pract. Conf., Nizhnevartovsk, December 08, 2021. Nizhnevartovsk : Nizhnevartovsk State University Publ. Pp. 284–288. DOI: 10.36906/AP-2022/47 (In Russian).
12. News dataset from Lenta.Ru. *Kaggle.com.* URL: https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta (accessed 12.04.2023).